



THE WORLD BANK



# Matching Techniques

Human Development  
Network

Middle East and North Africa  
Region

Spanish Impact Evaluation  
Fund

# The case of random assignment to the treatment

- If assignment to treatment is random in the population, both potential outcomes are independent of the treatment status

$$Y(1), Y(0) \perp D \quad (1)$$

- In this case the missing information does not create problems because:

$$E\{Y_i(0)|D_i = 0\} = E\{Y_i(0)|D_i = 1\} = E\{Y_i(0)\} \quad (2)$$

$$E\{Y_i(1)|D_i = 0\} = E\{Y_i(1)|D_i = 1\} = E\{Y_i(1)\} \quad (3)$$

- Then,

$$\delta = E\{\Delta_i | D_i = 1\} \quad (4)$$

$$= E\{Y_i(1) - Y_i(0) | D_i = 1\}$$

$$= E\{Y_i(1)|D_i = 1\} - E\{Y_i(0) | D_i = 1\}$$

$$= E\{Y_i(1)|D_i = 1\} - E\{Y_i(0)|D_i = 0\}$$

$$= E\{Y_i|D_i = 1\} - E\{Y_i|D_i = 0\}.$$

# The case of random assignment to the treatment



- Randomization ensures that the sample selection bias is zero:

$$E\{Y_i(0) \mid D_i = 1\} - E\{Y_i(0) \mid D_i = 0\} = 0 \quad (5)$$

- Note that randomization implies that the missing information is “missing completely at random” and for this reason it does not create problems.
- If randomization is not possible and natural experiments are not available we need to start from a different set of hypotheses.

# Unconfoundedness and selection on observables

- Let  $X$  denote a matrix in which each row is a vector of pre-treatment observable variables for individual  $i$ .

- **Definition** Unconfoundedness

*Assignment to treatment is unconfounded given pre-treatment variables if*

$$Y(1), Y(0) \perp D \mid X$$

- Note that assuming unconfoundedness is equivalent to say that:
  - within each cell defined by  $X$  treatment is random;
  - the selection into treatment depends only on the observables  $X$ .

# Average effects of treatment on the treated assuming unconfoundedness

- If we are willing to assume unconfoundedness:

$$E\{Y_i(0)|D_i = 0, X\} = E\{Y_i(0)|D_i = 1, X\} = E\{Y_i(0)|X\} \quad (6)$$

$$E\{Y_i(1)|D_i = 0, X\} = E\{Y_i(1)|D_i = 1, X\} = E\{Y_i(1)|X\} \quad (7)$$

- Using these expressions, we can define for each cell defined by  $X$ :

$$\begin{aligned} \delta_x &= E\{\Delta_i|X\} & (8) \\ &= E\{Y_i(1) - Y_i(0)|X\} \\ &= E\{Y_i(1)|X\} - E\{Y_i(0)|X\} \\ &= E\{Y_i(1)|D_i = 1, X\} - E\{Y_i(0)|D_i = 0, X\} \\ &= E\{Y_i|D_i = 1, X\} - E\{Y_i|D_i = 0, X\} \end{aligned}$$

# Average effects of treatment on the treated assuming unconfoundedness

- Using the Law of Iterated expectations, the average effect of treatment on the treated is given by:

$$\begin{aligned}\delta &= E\{\Delta_i | D_i = 1\} \\ &= E\{E\{\Delta_i | D_i = 1, X\} | D_i = 1\} \\ &= E\{E\{Y_i | D_i = 1, X\} - E\{Y_i | D_i = 0, X\} | D_i = 1\} \\ &= E\{\delta_x | D_i = 1\}\end{aligned}\tag{9}$$

- where the outer expectation is over the distribution of  $X | D_i = 1$ .

# Matching and regression strategies for the estimation of average causal effects

- Unconfoundedness suggests the following strategy for the estimation of the average treatment effect defined in equations 8 and 9:
  - i. stratify the data into cells defined by each particular value of  $X$ ;
  - ii. within each cell (i.e. conditioning on  $X$ ) compute the difference between the average outcomes of the treated and the controls;
  - iii. average these differences with respect to the distribution of  $X_i$  in the population of treated units.
  
- This strategy raises the following questions:
  - Is this strategy different from the estimation of a linear regression of  $Y$  on  $D$  controlling non parametrically for the full set of main effects and interactions of the covariates  $X$ ?
  - Is this strategy feasible?

# Is matching feasible? the dimensionality problem



- It is evident, however, that the inclusion in a regression of a full set of nonparametric interactions between all the observable variables may not be feasible when the sample is small, the set of covariates is large and many of them are multivalued, or, worse, continue.
- This dimensionality problem is likely to jeopardize also the matching strategy described by equations 8 and 9:
  - With  $K$  binary variables the number of cells is  $2^K$  and grows exponentially with  $K$ .
  - The number of cell increases further if some variables in  $X$  take more
  - than two values.
  - If the number of cells is very large with respect to the size of the sample it is very easy to encounter situations in which there are:
    - cells containing only treated and/or
    - cells containing only controls.

# Are matching and regression feasible: the dimensionality problem



- Hence, the average treatment effect for these cells cannot be computed.
- Rosenbaum and Rubin (1983) propose an equivalent and feasible estimation strategy based on the concept of *Propensity Score* and on its properties which allow to reduce the dimensionality problem.
- It is important to realize that regression with a not saturated model is not a solution and may lead to seriously misleading conclusions.

# Matching based on the Propensity Score

- **Definition** Propensity Score (Rosenbaum and Rubin, 1983):  
*The propensity score is the conditional probability of receiving the treatment given the pre-treatment variables:*

$$p(X) \Pr\{D = 1|X\} = E\{D|X\} \quad (10)$$

- The propensity score has two important properties:
  - **Lemma 1** Balancing of pre-treatment variables given the propensity score (Rosenbaum and Rubin, 1983)

*If  $p(X)$  is the propensity score*

$$D \perp X | p(X) \quad (11)$$

- **Lemma 2** Unconfoundedness given the propensity score (Rosenbaum and Rubin, 1983)

*Suppose that assignment to treatment is unconfounded, i.e.*

$$Y(1), Y(0) \perp D | X$$

*Then assignment to treatment is unconfounded given the propensity score, i.e.*

$$Y(1), Y(0) \perp D | p(X) \quad (12)$$

# Average effects of treatment and the propensity score

- Using the propensity score and its properties we can now match cases and controls on the basis of the propensity score instead of the multidimensional vector of observables  $X$ .

$$E\{Y_i(0)|D_i = 0, p(X_i)\} = E\{Y_i(0)|D_i = 1, p(X_i)\} = E\{Y_i(0)|p(X_i)\} \quad (13)$$

$$E\{Y_i(1)|D_i = 0, p(X_i)\} = E\{Y_i(1)|D_i = 1, p(X_i)\} = E\{Y_i(1)|p(X_i)\} \quad (14)$$

- Using these expressions, we can define for each cell defined by  $p(X)$ :

$$\begin{aligned} \delta_{p(x)} &\equiv E\{\Delta_i|p(X_i)\} & (15) \\ &\equiv E\{Y_i(1) - Y_i(0)|p(X_i)\} \\ &\equiv E\{Y_i(1)|p(X_i)\} - E\{Y_i(0)|p(X_i)\} \\ &= E\{Y_i(1)|D_i = 1, p(X_i)\} - E\{Y_i(0)|D_i = 0, p(X_i)\} \\ &= E\{Y_i|D_i = 1, p(X_i)\} - E\{Y_i|D_i = 0, p(X_i)\}. \end{aligned}$$

# Average effects of treatment and the propensity score

- Using the Law of Iterated expectations, the average effect of treatment on the treated is given by:

$$\begin{aligned}\tau &= E\{\Delta_i | D_i = 1\} \quad (29) \\ &= E\{E\{\Delta_i | D_i = 1, p(X_i)\} | D_i = 1\} \\ &= E\{E\{Y_i(1) | D_i = 1, p(X_i)\} - E\{Y_i(0) | D_i = 0, \\ &\quad , p(X_i)\} | D_i = 1\} \\ &= E\{p(x) | D_i = 1\}\end{aligned}$$

- where the outer expectation is over the distribution of  $p(X_i) | D_i = 1$ .

# Implementation of the estimation strategy



- To implement the estimation strategy suggested by the propensity score and its properties two sequential steps are needed.

*i. Estimation of the propensity score*

This step is necessary because the “true” propensity score is unknown and therefore the propensity score has to be estimated.

*ii. Estimation of the average effect of treatment given the propensity score*

Ideally in this step, we would like to

- match cases and controls with exactly the same (estimated) propensity score;
- compute the effect of treatment for each value of the (estimated) propensity score (see equation 28).
- obtain the average of these conditional effects as in equation 29.

# Implementation of the estimation strategy



- This is unfeasible in practice because it is rare to find two units with exactly the same propensity score.
- There are, however, several alternative and feasible procedures to perform this step:
  - Stratification on the Score;
  - Nearest neighbor matching on the Score;
  - Radius matching on the Score;
  - Kernel matching on the Score;
  - Weighting on the basis of the Score.

# Estimation of the propensity score

- ❑ Apparently, the same dimensionality problem that prevents the estimation of treatment effects should also prevent the estimation of propensity scores.
- ❑ This is, however, not the case thanks to the *balancing property* of the propensity score (Lemma 1) according to which:
  - ❑ observations with the same propensity score have the same distribution of observable covariates independently of treatment status;
  - ❑ for given propensity score assignment to treatment is random and therefore created and control units are on average observationally identical.

# Estimation of the propensity score

- Hence, any standard probability model can be used to estimate the propensity score, e.g. a logit model:

$$Pr\{D_i | X_i\} = \frac{e^{\lambda h(X_i)}}{1 + e^{\lambda h(X_i)}} \quad (16)$$

where  $h(X_i)$  is a function of covariates with linear and higher order terms.

- The choice of which higher order terms to include is determined solely by the need to obtain an estimate of the propensity score that satisfies the *balancing property*.

# Estimation of the propensity score



- Inasmuch as the specification of  $h(X_i)$  which satisfies the *balancing property* is more parsimonious than the full set of interactions needed to match cases and controls on the basis of observables (as in equations 8 and 9), the propensity score reduces the dimensionality of the estimation problem.
- Note that, given this purpose, the estimation of the propensity scores does not need a behavioral interpretation.

# An algorithm for estimating the propensity score



- i. Start with a parsimonious logit or probit function to estimate the score.
- ii. Sort the data according to the estimated propensity score (from lowest to highest).
- iii. Stratify all observations in blocks such that in each block the estimated propensity scores for the treated and the controls are not statistically different:
  - (a) start with five blocks of equal score range  $\{0 - 0.2, \dots, 0.8 - 1\}$ ;
  - (b) test whether the means of the scores for the treated and the controls are statistically different in each block;
  - (c) if yes, increase the number of blocks and test again;
  - (d) if no, go to next step.

(it continues in the next slide...)

# An algorithm for estimating the propensity score



- vi. Test that the *balancing property* holds in all blocks for all covariates:
  - (a) for each covariate, test whether the means (and possibly higher order moments) for the treated and for the controls are statistically different in all blocks;
  - (b) if one covariate is not balanced in one block, split the block and test again within each finer block;
  - (c) if one covariate is not balanced in all blocks, modify the logit estimation of the propensity score adding more interaction and higher order terms and then test again.
- Note that in all this procedure the outcome has no role.
  - See the STATA program `pscore.ado` downloadable at <http://www.iue.it/Personal/Ichino/Welcome.html>

# Some useful diagnostic tools

---

- ❑ As we argued at the beginning of this section, propensity score methods are based on the idea that the estimation of treatment effects requires a careful matching of cases and controls.
- ❑ If cases and controls are very different in terms of observables this matching is not sufficiently close and reliable or it may even be impossible.
- ❑ The comparison of the estimated propensity scores across treated and controls provides a useful diagnostic tool to evaluate how similar are cases and controls, and therefore how reliable is the estimation strategy.

# Some useful diagnostic tools

- More precisely, it is advisable to:
  - count how many controls have a propensity score lower than the minimum or higher than the maximum of the propensity scores of the treated.
    - Ideally we would like that the range of variation of propensity scores is the same in the two groups.
  - generate histograms of the estimated propensity scores for the treated and the controls with bins corresponding to the strata constructed for the estimation of propensity scores.
    - Ideally we would like an equal frequency of treated and control in each bin.
 

Note that these fundamental diagnostic indicators are not computed in standard regression analysis, although they would be useful for this analysis as well. (See Dehejia and Wahba, 1999).

# Comments and extensions

- *Irrelevant controls*

If the goal is to estimate the effect of treatment on the treated the procedure should be applied after having discarded all the controls with a propensity score higher than the maximum or lower than the minimum of the propensity scores of the treated.

- *Penalty for unequal number of treated and controls in a block*

- Note that if there is a block in which the number of controls is smaller than the number of treated, the variance increases and the penalty is larger the larger the fraction of treated in that block. If  $N_q^T = N_q^C$  the variance simplifies to:

$$Var(\tau^S) = \frac{1}{N^T} [Var(Y_i^T) + Var(Y_j^T)] \quad (20)$$

# Comments and extensions

---

- *Alternatives for the estimation of average outcomes within blocks*

In the expressions above, the outcome in case of treatment in a block has been estimated as the average outcome of the treated in that block (and similarly for controls).

Another possibility is to obtain these outcomes as predicted values from the estimation of linear (or more sophisticated) functions of propensity scores.

The gains from using these more sophisticated techniques do not appear to be large. (See Dehejia and Wahba, 1996.)

# Estimation of the treatment effect by Nearest Neighbor, Radius and Kernel Matching



- ❑ Ideally, we would like to match each treated unit with a control unit having exactly the same propensity score and viceversa.
- ❑ This exact matching is, however, impossible in most applications.
- ❑ The closest we can get to an exact matching is to match each treated unit with the *nearest* control in terms of propensity score.
- ❑ This raises however the issue of what to do with the units for which the nearest match has already been used.
- ❑ We describe here three methods aimed at solving this problem.
  - Nearest neighbor matching with replacement;
  - Radius matching with replacement;
  - Kernel matching

# References

---

- Dehejia, R.H. and S. Wahba (1999), “Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs”, *Journal of the American Statistical Association*, 94, 448, 1053-1062.
- Dehejia, R.H. and S. Wahba (1996), “Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs”, Harvard University, Mimeo.
- Hahn, Jinyong (1998), “ ON the role of the propensity score in efficient semiparametric estimation of average treatment effects”, *Econometrica*, 66,2,315-331.

# References

---

- Heckman, James J. H. Ichimura, and P. Todd (1998), “ Matching as an econometric evaluation estimator ”, *Review of Economic Studies*, 65, 261-294.
- Hirano, K., G.W. Imbens and G. Ridder (2000), “Efficient Estimation of Average Treatment Effects using the Estimated Propensity Score”, mimeo.
- Rosenbaum, P.R. and D.B. Rubin (1983), “The Central Role of the Propensity Score in Observational Studies for Causal Effects”, *Biometrika* 70, 1, 41–55.